ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

单篇论文被引频次影响因素及预测研究综述

张素芳 刘慧敏
华南师范大学经济与管理学院 广州 511400

摘要：[目的 / 意义] 梳理单篇论文被引频次的相关影响因素以及被引频次预测研究现状，为科研人员和科研机构研究单篇论文被引频次影响因素及预测提供一个全面系统的认知框架。[过程 / 方法] 采用文献调研法，通过对现有文献进行系统的梳理，总结被引频次预测的影响因素、研究对象和研究方法的相关内容和特点，并通过列表的方式对比分析不同的方法，总结现有研究普遍存在的问题和一些创新的解决方案。[结果 / 结论] 在系统梳理和总结的过程中发现，影响因素与预测结果之间因果关系不明确，研究样本数据缺乏多样性，未明确研究结果的适用性与预测周期的关系，模型评估可解释性较弱。因此，应从解决问题的前提条件、选择有针对性的样本、改进影响因素提取方法、运用数学思维方式进行建模等方面提高后续研究的质量。

关键词：被引频次预测　影响因素　回归分析　机器学习　深度学习

分类号：G251

引用格式：张素芳 , 刘慧敏 . 单篇论文被引频次影响因素及预测研究综述 [J/OL]. 知识管理论坛 , 2022, 7(3): 299-313[引用日期]. http://www.kmf.ac.cn/p/294/.

1 引言

科学系统包含了大量元素和链接，研究者对学术论文的引文动态和科学演变越来越感兴趣。被引频次在一定程度上反映了论文受到的关注程度，然而通常只有少数的研究论文积累了绝大多数的被引频次，而其他大多数论文只吸引了少数的其它论文的引用 [1]。也就是说，一些研究论文比其他研究论文更有可能吸引研究者的注意。对于不断增长的文献数量，预测哪篇论文更有可能引起学术界的关注是很重要的。因此，被引频次预测成为目前文献计量领域的一个新的研究方向。该研究主题已经涌现了不少的论文，在研究建模过程中，一些研究人员被大量的低被引频次的论文所困扰，方法和影响因素特征的选择多样化，导致研究的重复累赘，尽管已经有学者对该主题进行系统性的综述，但是主要集中在影响因素和研究方法上，还未有学者从研究人员如何介入该领域研究提出有效的解决方案。基于此，本文梳理了论文被引频次的影响因素，面向预测任务，将

作者简介：张素芳，副教授，博士，研究生导师；刘慧敏，硕士研究生，通信作者，E-mail: 1273569816@qq.com。

收稿日期：2021-12-01　　发表日期：2022-06-15　　本文责任编辑：刘远颖

被引频次预测分为回归任务和分类任务，阐述这两个方面单篇论文被引频次的研究方法、论文的研究对象形式和预测周期等，最后根据现有研究中的普遍问题提出一些方案，以期为后续研究者提供借鉴和参考。本文主要的梳理框架如图 1 所示：

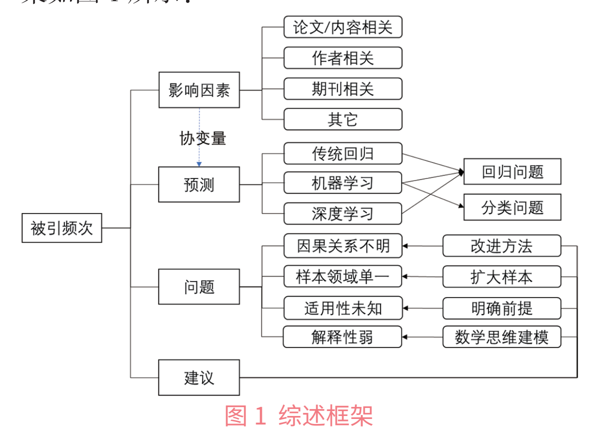


图 1 综述框架

2 单篇论文被引频次影响因素

学术论文的被引频次预测已经被广泛地研究，在这些被引频次预测的研究中，研究人员往往关注什么因素会影响论文的被引量，从而筛选重要的影响因素来对引文的被引量进行预测。F. Didegah 和 M. Thelwall[2] 认为，论文引用动机复杂，引用者对论文的智力认知是论文被引量的内在因素，其可以通过访谈和问卷进行调查，但是其具有耗时的缺点，并且由于引用动机的复杂性和学科依赖性，这种定性研究通常只涉及一小部分学者样本，而外部因素可以大规模地量化和计算，因此可以用来预测未来的引文影响。影响被引率的外在因素包括被引用论文的作者、摘要、期刊、领域和参考文献以及论文本身等属性特征。本文研究仅局限于外部动机，将这些因素归纳为论文本身、作者、期刊、其他四大类。

2.1 与论文本身相关的影响因素

在与论文相关的影响因素中，与被引频次相关的主要因素之一是论文的主题，论文的主题是论文研究内容的核心，它可以用来预测论文未来被引频次 [3]。论文的内容可以从三个维度进行评价——论文所研究的主题的关注度、主题新颖性、主题的多样性。热门的主题通常也会吸引更多的关注和更多其它论文的引用 [4]，论文主题新颖也会增强其影响力和被引率 [5]，论文主题越有吸引力和新颖性越高，它被引频次可能就会越多。此外，所研究的主题范围和主题领域将影响被引频次，论文研究主题的多样性会给论文的被引频次带来影响 [6]。

在主题的识别研究中，大多数研究者都是使用隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）模型或其衍生模型进行主题识别，进而计算其主题的关注度 / 热度、新颖性、多样性等指标。主题关注度的测度主要从累计被引的角度进行计算，多样性的测度主要从信息熵的角度进行计算，新颖性的测度主要从同行评议、引用、内容三个角度进行计算 [7]，除去同行评议，另外两种方法（引用对的共现频率和主题内容的共现频率）都是基于一种共现思想进行考虑的。关于内容新颖程度的研究有许多，但其研究的角度大致相同。

参考文献的数量、权威度以及论文中参考文献的多样性也会增加论文的被引频次 [8]。参考文献数量多的研究与较高的被引率相关 [9]。平均参考文献年份越年轻的论文，可能获得更多的被引量，引用“旧出版物”的论文被引量明显减少 [10]，因为论文信息随着时间的流逝会过时 [11]。一般来说，在发表后的前几年，被引频次达到峰值，随着时间的推移，被引频次逐渐减少。此外，参考文献的权威度（累计被引频次 [12]）和多样性（施引文献所属研究领域 [6] 和跨国籍 [2]）也会对论文引文率产生影响。

还有研究发现，某些类型的文档比其他类型的文档获得更多的被引，如综述论文比研究论文的被引用更多 [13]。基金资助是科学研究的重要经济来源，充足的经费可以使研究得到较好的物质保障，一般来说，获得更高水平资助的研究项目的论文能比未受资助的论文获得更多的被引 [8]。在一些研究中，论文早期被引率及其速度也被认为是其未来被引的预测因素 [6]。

知识管理论坛，2022（3）:299-313
DOI: 10.13266/j.issn.2095-5472.2022.025

论文早期被引是科学界对这篇论文的早期反馈，其被引速率在一定程度上反映了论文在科学界的传播速度。论文的长度（其页数）也是增加被引频次的因素之一[8]，因为较长的论文包含了更多的信息[14]。论文的标题是整篇论文内容最浓缩的概括，也是研究学者搜索论文最先看到的内容，因此也有研究学者对这方面的内容进行了研究，H. R. Jamali 和 M. Nikzad[15] 认为，一个信息丰富的标题可以增加论文的被引次数，但标题长度和被引之间没有显著的相关性，相对于被引频次，标题特征对下载次数的影响更大[16]。开放获取是指论文的可访问性和可见性，能够阅读全文作者才能从该篇论文中获取自己需要引用的内容，因此发表在开放获取期刊上的论文，往往比发表在非开放获取期刊上的论文被引量更多[817]。

与论文相关的影响因素研究范围非常广泛，除了以上研究得比较多的影响因素，国外学者还对论文的方法论 / 研究设计、章节特征、是否使用数据 / 附录等方面进行了详细的研究[18]。尽管有论文研究这些因素在某些领域上与被引率是有相关性的，但在不同领域的研究中，这些因素与被引率不一定产生关系，或者只有微弱的关系。这些研究经常忽略不同学科的差异[19]，其实一些影响因素都是具有明显的学科领域特征的，因此，构建普适性的综合指标并不是很好的选择。笔者对以上综述的影响因素进行了整体归纳，如表 1 所示：

表 1 论文相关影响因素及描述

影响因素		描述
研究主题	主题新颖性	某主题在论文集合中的新颖程度
	主题关注度	其他科研工作者对论文主题的关注程度
	主题多样性	论文研究主题个数
参考文献	数量	参考文献的篇数
	累计被引	截止计算日期参考文献的累计被引
	多样性	施引文献所属研究领域；跨国籍
	平均参考年份	所有参考文献出版年份的平均值
论文类型		综述型论文、研究型论文
基金资助		是否有基金资助，基金资助类型
早期被引和被引速率		论文早期收到的被引频次；以及早期论文被引的速度
开放获取		论文的可访问性和可见性
标题特征		标题长度、有无标点符号、标题类型（复合标题、问题标题、描述性标题）
论文长度/篇幅		以页数表示长度
方法论/研究设计		方法学的质量；方法类型；方法描述
章节特征		章节内容
数据/附录		用数据呈现结果，论文带有附录

2.2 作者相关影响因素

与作者相关的因素也会对论文的被引频次产生影响。作者的数量是一种表明研究合作程度的衡量标准。高质量的论文往往涉及多个科研人员的合作，作者合著（特别是国际上的合作[20]）能增加论文的被引率[21]。然而，有一些研究却发现了相反的结果，证明国际合作与论文被引频次之间并无特别的联系[8]。随着引文

时间窗口变长，作者数量与引文之间的相关性也会减弱[22]。但也有研究报告指出，不同领域的作者合作能增加论文的被引率[23]。因此，作者之间的合作是否影响论文的被引频次存在着较大的分歧。

此外，论文作者数量和自引数呈正比的关系[24]，但是，有研究发现，自引率与非自引率的比例随着论文积累的总被引频次的增加而降低，自引往往集中出现在论文出版后的很短的时间内[25]。因此从宏观角度看，在分析论文被引时不需要在分析中排除自引[26]。

著名作者在其研究领域有着较高的声望，其论文往往会有较高的被引量[27]。马太效应使具有高被引特征的作者发表的论文比具有低被引特征的作者发表的论文更能获得其它论文的引用[12]。因此，作者之前的论文的被引频次可以被认为是对未来论文被引的一个很好的预测因素[28]。H 指数是用来衡量科学界研究者能力的最常用的标准[29]，声望高的作者 H 指数往往很高，因此，在研究作者某一个领域的声望对论文被引量的影响时，常用 H 指数作为一个计量的指标。作者所属机构的声望很大部分依赖于作者。一般来说，排名高的学校的论文会有更多的被引量[30]。

除此之外，关于作者的人口统计学特征也被纳入到测量指标之中。有研究发现，白人和男性比非白人和女性有更高的影响力[31]。但也有研究表明人口统计学特征对于论文是否被引并无显著性的影响[32]。

笔者对作者相关的影响因素做了以下的归纳，如表 2 所示：

表 2 作者相关影响因素及描述

影响因素	描述
数量	参与论文撰写的人数
合作方式	国际合作、国内合作、组织内合作、组织外合作
H指数	作者发表的N_p篇论文中有h篇每篇至少被引h次
累计被引量	截至计算日期，作者获得的累计被引量
自引率	论文作者对论文的被引量占比
作者所属机构	作者撰写论文时所属单位
人口统计学	性别、年龄、种族、国家

2.3 期刊相关影响因素

除了与论文与作者相关方面的影响外，有研究发现论文的被引频次的主要决定因素是期刊层面的因素[33]。论文在出版期刊上获得的平均被引量可以预测论文未来的被引量[6]。研究者在发表论文时往往会更倾向于发表在具有高影响力的刊物上，以提高他们论文的可见性，从而获得更高的被引量。研究证明，在具有高影响力的期刊上发表论文能比在低影响力的期刊上发表的论文更容易获得高被引[34]。尽管大量研究都证明了期刊的影响力与论文的被引量存在正相关关系，但是也有一些研究发现，期刊影响因子不一定是被引频次预测的影响指标[35]。也有研究者使用出版物的总被引量、生产力（刊载论文数）作为研究的影响因素之一[36]。除此之外，部分研究认为期刊的语种对于论文被引率来说也是有一定的影响的[32]，特别是英语期刊[12]，会积累更多的被引量。以下是本文对期刊相关影响因素的归纳，如表 3 所示：

知识管理论坛，2022（3）:299-313
DOI: 10.13266/j.issn.2095-5472.2022.025

表 3 期刊相关影响因素及描述

影响因素	描述
影响因子	某时间段积累的平均被引量
总被引量	期刊上在某年度刊载论文的总被引量
生产力/发文量	刊载的论文数
语言	期刊的语言类型

2.4 其他影响因素

随着研究的不断深入，出现了社交网络、时间等因素等新的研究视角。研究者开始分析社会网络活动和文献计量学之间的潜在联系[37]。孔玲等[38]在归纳相关影响因素时，增加了替代计量角度的因素，但替代计量因素针对的是开放学术网络平台及社交网站进行研究，与传统的学术论文网站存在一定的区别。除了社交网络外，学术引文网络也是一个很重要的因素。为了衡量作者的社交性，R. Yan 等建立了一个作者协作网络，并用 PageRank 递归地计算了社交性[39]。由于学术论文的引用具有半衰期属性，所以时间因素对于论文的被引频次预测来说也是一个非常具有研究价值的因素。E. Butun 和 M. Kaya 将作者的引文网络和时间因素相结合，引入一个时间链路指标，考虑作者引文网络的演化趋势，利用复杂网络中的局部和全局拓扑结构，根据引文网络中的链路来预测链接的权重，这是第一个使用定向、加权和时间引文网络来进行被引频次预测的研究[40]。

笔者对其他类因素进行了归纳，如表 4 所示：

表 4 其他影响因素及描述

影响因素	描述
替代计量	社交媒体的转载、评论、收藏、下载的行为
学术网络	论文之间、作者之间构成的引文网络
时间	赋予时间权重

3 单篇论文被引频次预测方法

随着科学计量的发展，众多的研究方法被引进到被引频次预测研究中。从任务导向出发，可以将预测问题定义为回归问题，也可以将预测问题定义为分类问题。回归问题中，主要的研究方法分为以下三类：传统的回归分析方法、机器学习方法、深度学习方法。而分类问题，则主要是使用机器学习的方法进行研究。在引入的多种方法中，每种研究方法都有其特性和适用性。

3.2 定义为回归问题的预测方法

将被引频次预测定义为回归问题，是指利用一篇论文的相关特征，预测这篇论文在某个时间节点的被引频次[41]。回归是目前最常用的一种预测方法[42]。本文将从传统的回归方法、机器学习方法和深度学习方法三个方面梳理论文的被引频次预测研究现状。

3.1.1 传统回归预测

在预测回归问题上，早期研究人员更多地使用传统的线性回归方法进行拟合研究，C. Lokker 等[43]人使用了 17 个参考文献相关特征和 3 个期刊相关特征来预测临床论文两年被引频次，其多元回归预测结果训练集的决定系数 r^2 为 0.60，测试集的决定系数 r^2 为 0.56，在进行被引频次预测敏感度分析时，被引频次排名前半部分和前三分之一的论文特异性为 72% 和 82%，回归预测对于高被引论文的预测效果更好，该结论并不仅仅在这篇文章中得到体现，G. Abramo 等[44]的研究中也有提及，其反映的事实是绝大多数论文是低被引的，只有少部分论文是高被引的[36]。T. Yu 等[28]采用多元逐步回归的方法，从论文的外部特征、作者的特征、发表期刊的特征和被引论文的特征中选择好的特征变量，建立一个描述特征与引文影响之间关系的模型，用于预测论文发表 5 年后的被引频次。L. Bornmann 等[45]使用了 WoS 数据库中 1980 年发表的所有论文，涵盖各个学科，总计约 50 万篇学科文献，以发表后的第 31 年被引频次作为因变量，进行论文的长期影响预测，研究发现，只有论文发表后前几年的被引频次能显著提高论文的长期影响预测，同样的

研究结果也被 G. Abramo 等发现。G. Abramo 等[44] 使用了两种线性回归模型，预测的平均准确性对于两年以上的引文时间窗口是良好的，三年的引文时间窗口足够预测科学文献的长期影响，该模型对于低被引的科学文献预测准确率较低，并且不同学科的准确率也不同。程子轩等[46] 使用逐步回归的方法，对图书情报期刊论文发表后的第七年被引频次进行预测，实验发现了 10 个与学术论文被引频次呈显著相关的影响因素。

传统的回归分析方法是基于统计学进行的，这类模型对于小数据量、简单的关系很有效，并且有直观的理解和解释，但是对于数据分布的要求十分高，对于结构复杂的数据其处理精准度并不是很高。传统回归方法预测论文被引频次的部分论文如表 5 所示：

表 5 传统回归方法预测论文被引频次的部分论文（回归问题）

序号	研究对象	影响因素	研究方法/工具	是否包括冷启动	预测	论文来源
1	WoS临床领域文章	文章特征、期刊特征	多元回归	是（发表后3周内数据）	发表后两年内的被引频次	[43]
2	Thomson ISI信息科学与图书馆科学期刊	论文的外部特征、作者特征、引用特征、期刊特征	多元回归	否（前2年的被引频次）	发表5年后的被引频次	[28]
3	WoS1980年发表的所有论文	作者数量、被引文献数量和页数等	最小二乘回归	否（1-30年的被引频次百分位）	发表后的第31年被引频次百分位数	[45]
4	WoS有关意大利的出版语料库	早期引文、IF（期刊影响因子）	线性回归模型	否（0-8年的早期被引）	发表文章9年后的被引频次	[44]
5	CNKI图书情报期刊论文	作者特征、期刊特征	逐步回归	是	发表后第7年的被引频次	[46]

3.1.2 机器学习预测

随着科学技术的发展，机器学习开始出现在被引频次预测研究中，R. Yan 等[47] 利用高被引论文的基本特征，使用了多种机器学习方法进行比较，预测每种文献的被引频次，其最佳预测模型 CART 分类回归树在预测 10 年内的被引频次其决定系数 r^2 平均预测性能为 0.786，其研究发现，作者的专业知识和期刊的影响力是该研究的显著影响因素，孤立的内容特征无法进行被引频次预测。T. Chakraborty 等[6] 则认为，大多数的回归方法存在一个潜藏的假设，即所有发表论文的引文模式都具有相似的特征，该假设在一定程度上影响了预测的准确性，为此，他提出使用分层学习的方法，将论文分为了 6 种引文模式，分别对不同模式的论文使用支持向量机模型进行回归模拟，其研究证明，分层学习是有效的，但该方法仅对于平均每年被引频次大于 1 的论文有效。J. Chen 和 C. Zhang 基于 6 种内容特征和 10 项作者特征，引入 IBM 模型提取内容特征计算论文主题之间的关联概率，并使用二部网络投影得到作者协作网络，使用梯度增强回归树 (GBRT) 来预测论文的引文计数，实验结果表明，GBRT 的“内容特征”组在 KDDCUP 数据集上的性能最高[48]。然而，在 X. Zhu 和 Z. Ban[36] 的研究中，其使用 ArnetMiner 数据集，引入学术网络特征进行研究，发现作者的特征更重要，支持向量机 SVM 的 r^2 最高，达到 88.87%。机器学习方法预测论文被引频次的部分论文见表 6。

3.1.3 深度学习预测

最近几年，神经网络等深度学习方法开始被应用于被引频次预测。深度学习模型是一种特殊的机器学习，它允许模型通过多个处理层学习具有多个抽象层次的数据[49]。在深度学习中，RNN、LSTM、GRU 等时间序列神经网络可以预测未来一段时间的序列值，BP 神经网络和 CNN 对于特征值处理更加有效。

ChinaXiv合作期刊
知识管理论坛，2022（3）:299-313
DOI: 10.13266/j.issn.2095-5472.2022.025

表 6 机器学习方法预测论文被引频次的部分论文（回归问题）

序号	研究对象	影响因素	研究方法 / 工具	是否包括冷启动	预测	论文来源
1	ArnetMiner 学术数据集（计算机领域）	内容特征、作者特征、期刊特征	GPR、CART、KNN、LR、SVR	是	1、5、10 年内的被引频次	文献 [47]
2	从微软学术搜索 (MAS) 中抓取的公开数据集（计算机科学领域）	引用模式	支持向量机	是	发表 1-5 年后的被引频次	文献 [6]
3	KDD CUP 数据集（高能物理理论）	内容特征、作者特征	IBM 模型、梯度增强回归树（GBRT）	否（第一年的早期被引）	3 年后的被引频次	文献 [48]
4	ArnetMiner 学术数据集（计算机领域）	学术网络特征	GPR、DNN、MLR、SVM	是	发表 3 年后和 5 年后的被引频次	文献 [36]

A. Abrishami 等 [50] 利用 RNN 循环神经网络学习论文的引文序列从而预测未来引文序列，但是在进行预测过程中，仅仅使用了论文发表后早期引文特征，并未将其他信息源如作者的功能、论文的内容等作为数据进行输入。LSTM 模型是 RNN 模型的变种，S. Yuan 等 [51] 结合了论文的内在质量、老化效应、马太效应和近期效应 4 种现象，提出了基于 RNN 和 LSTM 的论文被引频次预测模型，但也仅是使用时间序列进行预测，未使用作者、期刊、论文等相关特征。与前文多提到的研究相比，J. Wen 等 [52] 则提取了用于预测论文被引频次的特征，然后将这些特征输入到 GRU 神经网络中进行预测。将预测结果与其他回归模型进行了比较。实验结果表明，该模型预测精度高，收敛速度快。引文计数的时间序列预测优于现有的方法。

区别于时间序列数据预测方法，X. Ruan 等 [42] 使用四层反向传播（BP）神经网络模型来预测论文未来某个时间段总被引频次，其研究结果发现，BP 神经网络的性能明显优于 6 个基线模型（XGBoost、RF、LR、SVR、KNN、RNN）。在预测效果方面，低被引论文的准确率高于高被引论文。J. Xu 等 [53] 则提出了一种以数据为中心的方法，结合许多文献特征，使用卷积神经网络（CNN）来预测长期的科学影响。

与依赖于统计学的线性回归模型不同，深度学习方法对实验数据的分布没有严格的要求。神经网络的预测结果通常是具有鲁棒性的。此外，浅层机器学习模型的性能取决于特征工程的质量，特征工程质量越好，模型的学习效率往往会越高。然而，特征工程的构建、选择和提取并非易事。相比之下，深度神经网络在其特征学习方面具有优势——自动特征工程 [49]，即它可以通过多层次和非线性变换，将初始的“底部”特征表示自动转换为“高级特征” [42]。深度学习方法预测论文被引频次的部分论文如表 7 所示。

3.1.4 小结

上述提及的预测研究大部分都有对论文进行筛选处理，即删除低被引论文后，再进行预测。其原因是低被引论文在回归预测上的效果并不明显，回归预测在很多情况下仅适合预测高被引论文，然而，一篇新出版的论文，我们并不知道其是否属于高被引论文，因此预测效果与实际应用会产生较大的差别。Y. DONG 等 [54] 认为被引频次预测具有长尾效应，不适合采用回归方式进行预测，即预测的有效性从根本上受到被引频次的幂律分布的限制，低被引论文普遍存在，而高被引论文则相对罕见。由于绝大多数文献积累的被引频次很少，传统的回归分析将很难度量论文的被引频次。为了解决这种困难，通过提取高被引论文的特征，并将这些特征映射到论文的被引频次上，可以一定程度

上提高被引频次的预测效率，但是由于低被引论文的数量太多，导致高被引论文的特征并不非常明显，这将会使得实际应用数据集的预测效果大大降低。

表 7 深度学习方法预测论文被引频次的部分论文（回归问题）

序号	研究对象	影响因素	研究方法 / 工具	是否包括冷启动	预测	论文来源
1	WoS 中 5 种期刊（《自然》《科学》《新英格兰医学杂志》《细胞》）和《美国国家科学院院刊》	发表后 5 年的被引频次序列	RNN 神经网络	否	发表后 14 年的被引频次序列	文献 [50]
2	学术挖掘和搜索平台 AMiner 数据集	发表后 5 年的被引频次序列	LSTM 神经网络	否（发表前 5 年的被引频次序列）	未来 1-5 年的被引频次	文献 [51]
3	学术挖掘和搜索平台 AMiner 数据集	作者特征；期刊特征；论文特征	GRU-CPM 神经网络	否（发表前 5 年的被引频次序列）	未来 1-5 年的被引频次	文献 [52]
4	2000 年至 2013 年在 CSSCI 图书馆、信息和文献领域发表的评论和研究论文	内容特征、作者特征、期刊特征、参考文献特征、早期引用特征	BP 神经网络	否（前 2 年被引频次）	未来 5 年的被引频次	文献 [42]
5	学术挖掘和搜索平台 AMiner 数据集	时间异构网络特征	CNN 卷积神经网络	是	发表后 10 年的被引频次	文献 [53]

3.2 定义为分类问题的预测方法

被引频次预测问题从回归转化为分类问题，尽管预测粒度变粗，但是预测结果更加符合引文数据分布规律，使得模型更加具有泛化性[41]。相比于回归问题的预测方法，将预测任务视为分类问题的研究方法则比较单一，主要是使用各种机器学习的方法进行分类预测。由于分类任务是有监督的学习，因此这类研究方法需要设定一个分类阈值，用以确定每篇文献的标签。常用来进行论文被引频次预测的分类方法有支持向量机（SVM）、贝叶斯网络（NB）、K 最近邻 (KNN)、逻辑回归（LRC）、决策树、梯度提升决策树（GBRT）、袋装法（BAG）、随机森林（RF）、XGBoost、AdaBoost 算法等。

A. Ibanez 等 [55] 将论文分为三类——很少被引（被引频次小于等于 1）、一些被引（被引频次 2-4）和许多被引（被引频次超过 4），采用机器学习方法，如朴素贝叶斯、逻辑回归、决策树和 k 最近邻 (KNN)，来预测从第一年到第四年的被引频次，结果表明，逻辑回归算法和朴素贝叶斯算法的准确率最高。L. Fu 和 C. Aliferis[4] 使用支持向量机 (SVM) 在生物医学领域预测一篇论文发表 10 年后被引量是否高于某个阈值（20、50、100、500），模型的预测 AUC（Area Under Curve，线下曲线面积）为 0.857-0.918。M. Wang 等 [56] 将天文学和天体物理学领域的 219 篇论文分为高、中、低三组，使用了一个由 5 个决策树分类器组成的多分类器系统来进行预测，并获得了较高的分类能力，其研究表明，论文的内部质量和外部特征（主要表现为作者和期刊的声誉），有助于提高论文的被引频次预测。Y. Dong 等 [54] 的研究发现，作者出版文献的主题和刊载期刊决定一篇论文是否将贡献其主要作者的 h 指数，发表文献的主题受欢迎程度和合著者的影响与预测目标无关，在预测一篇论文是否会在 5 年内对其主要作者的 h 指数有贡献时，其最佳模型具有 87.5% 以上的准确度。耿骞等 [41] 通过大量实验分析发

ChinaXiv合作期刊
知识管理论坛，2022（3）:299-313
DOI: 10.13266/j.issn.2095-5472.2022.025

现 GBDT、XGBoost 和随机森林的预测能力较强，且预测的时间段越长，效果也就相对越好。

机器学习的方法在识别高影响力或高被引论文上具有较高的准确度。但是，分类模型的分类标准并没有进行统一的界定，往往是研究人员根据所使用的论文数据集进行自定义界定，甚至同一研究人员在不同研究时期的分类标准都不一样，显示出分类方法具有粗粒度的缺点，该缺点限制了论文研究成果的普及应用 [42]；其次，分类结果是某一段时间内的被引总量，是论文被引量的简化处理 [42]，因此无法判断论文随时间变化而产生的被引趋势变化。

机器学习可以处理两类预测问题，即回归问题和分类问题。在众多的研究中，集成的机器学习方法和支持向量机都有比较好的预测效果。相比于预测回归值，机器学习在分类回归上有更好的表现。尽管分类预测粒度较粗，但是更能符合实际的应用数据，可以减少低被引数据在分类过程中的影响。机器学习方法预测论文被引频次的部分论文如表 8 所示：

表 8 机器学习方法预测论文被引频次的部分论文（分类问题）

序号	研究对象	影响因素	研究方法	是否包括冷启动	分类阈值	预测	论文来源
1	《生物信息学》	摘要	贝叶斯网络 (朴素贝叶斯和 K2)、逻辑回归、决策树、k 最近邻 (KNN)	是	few、some、many	发表后 1-4 年的被引频次	文献 [55]
2	《美国医学杂志》《内科医学年鉴》《英国医学杂志》《美国医学会杂志》《柳叶刀》《新英格兰医学杂志》	内容特征、文献计量学特征	支持向量机（SVM）、决策树	是	20、50、100、500	发表后 10 年内被引频次	文献 [4]
3	天文学和天体物理学领域论文 219 篇	论文外部特征；论文质量特征	决策树多分类器、遗传算法	否（论文发表前 5 年）	高（275 以上）、中（40-275）、低（40 以下）	—	文献 [56]
4	ArnetMiner 学术数据集	论文的作者、内容、发表地点和参考文献，以及与作者相关的社会和时间效应	逻辑回归 (LRC)、随机森林 (RF)、袋装决策树 (BAG)	是	主要作者的 H 指数	发表后 5 年内对主要作者的 h 指数作贡献	文献 [54]
5	WoS 的情报学和图书馆学 (Information Science & Library Science) 论文数据	期刊相关；作者相关；论文相关	朴素贝叶斯 (NB)、逻辑回归 (LR)、支持向量机 (SVM)、梯度提升决策树 (GBDT)、XGBoost、AdaBoost、随机森林 (RF)	是	作者在论文发表当年的篇均被引频次	发表后的 1 年、5 年、10 年	文献 [41]

4 被引频次预测研究中存在的一些共性问题分析

综合来看，不论是将预测研究定义为回归问题还是分类问题，在研究过程中都存在着一些共性的问题，本文将会对这些存在的问题进行分析。

4.1 影响因素与预测结果之间因果关系不明确

影响因素与被引频次之间更多的是相关性研究，两者之间是相关的并不意味着在预测模型中有较好的效果。由于被引频次相关的影响因素众多，关于被引频次预测的影响因素研究

已有较多的成果，各方面的影响因素均有涉及与研究，总体来说主要是论文 / 内容相关的影响因素、作者相关的影响因素、期刊相关的影响因素，还有一些其他影响因素，包括但不限于时间因素、替代计量因素、网络特征因素等。但是不同的数据集中，不同的影响因素可能会产生不同的效果，如 KDDCUP 数据集中，J. Chen 和 C. Zhang 研究发现内容特征更重要 [48]，而在 ArnetMiner 数据集中，X. Zhu 和 Z. Ban 发现作者特征更加重要 [36]。

4.2 研究样本数据缺乏多样性

被引频次预测研究的样本数据相对单一，使用的数据集大多是关于理工科和医学类科学文献。尽管有些研究中，有进行学科之间的对比，但是学科领域并未跳脱自然科学和人文科学之间的界限，因此研究缺乏全面性。ArnetMiner 学术数据集和 AMiner 数据集是使用较多的关于计算机领域的科学文献公开数据集，此外生物医学类的数据集也比较多，人文社科类数据集非常少，并且使用的数据集大多数来源于外文数据库。这种现象值得我们思考，已有的研究发现，不同领域的研究数据集之间差异比较大，因此，将这些被引频次预测的研究方法迁移到国内数据集或人文社会数据集是否依然适用有待验证。

4.3 未明确研究结果的适用性与预测周期的关系

预测未来长期影响最终目的还是落实到应用中，但是大多数论文并未对多长的周期是适用的进行阐述。在以上众多研究中，预测的周期长短不一。它们的研究目的是预测论文的短期或长期影响，以未来一定时间段的被引频次来衡量，该时间段在不同的研究中设定不一，如 1 年、5 年、10 年甚至 31 年的长度等。不同研究者使用的数据不同，造成研究的周期不同，但在多数研究论文中，并未阐述论文所研究的周期有何依据。只有少数论文对整体数据进行了研究，再划分出有效的引文时间窗口。引文时间窗口又引申了一个实用性问题，过长的引文时间窗口会出现信息的滞后性，导致预测结果无效，过短的引文时间窗口可能会造成模型准确度下降。

4.4 模型评估可解释性较弱

被引频次预测需要一个评价标准来对模型的好坏进行评估，常用的评估方法有决定系数 r^2、均方误差 MSE、平均绝对误差 MAE、准确率 ACC 等，但是在许多研究中，仅给出了评估方法的值大小以判断模型的好坏，对值大小并未进行详细的解释，这是这类研究的通病。事实上，模型评估方法的值大小是基于实际值和预测值进行计算的，如 MAE 是平均绝对误差，在进行值大小判断时，应该与真实值的大小进行比较，看误差值在真实值多大范围内，而不仅仅是比较不同方法产生的误差值大小。

5 提高被引预测研究质量的建议

针对第 4 部分提出的被引预测研究中存在的共性问题，本文提出了一些建议，希望能够给相关研究人员提供一些参考，以提高研究的质量。

5.1 改进影响因素提取方法，增强影响作用的针对性

上述综述已经从各个方面综合阐述了影响论文的因子，这些影响因子最终都有可能成为建立模型的特征之一。但是如何使这些特征因子更能表达出模型所需要的信息，我们需要从微观的具体操作方法的角度进行创新和应用。

在提取高级语义特征学习引文时间序列的研究中 [57]，其研究的核心是从元数据文本中获取语义信息，使用 Doc2Vec 算法对元数据文本中的句子进行编码，然后进一步通过 Bi-LSTM 和注意机制从句子嵌入中提取高级（段落级）语义特征，最后通过整合早期的引文来学习引文预测任务。该研究证明元数据语义特征对提高被引预测性能是有用的，为引文预测提供了一种很有前途的方法。

与主题相关的特征研究也是基于文本内容（标题、摘要等文本内容）进行挖掘的，但该研究与之不同的地方在于特征挖掘的粒度不同。

知识管理论坛，2022（3）:299-313
DOI: 10.13266/j.issn.2095-5472.2022.025

主题特征描述的是整篇文档的特征，常用的提取方法是 LDA 及其改进模型，所形成的是在语料库中通过参数调整得到的数量一定的主题，粒度相对较粗，少部分论文不一定能找到相对合适的主题。而元数据语义特征在 Doc2Vec 算法的基础上，进一步使用 Bi-LSTM 和注意机制进行语义挖掘，其粒度相对较细，使每一篇论文都能找到其特定的语义特征。

5.2 扩大研究样本，预测限定模型的适用性

在被引频次预测研究中，大多数研究通常使用单一数据集，因此研究得出的结果并不都适用于其他数据集。已有研究也表明，不同研究领域的数据集之间被引频次预测差异较大，因此为了使研究结果更具有普遍性和泛化性，应该使用更加全面的数据集，对差异较大的领域进行比较研究，分析影响不同预测结果的原因，使得研究更加严谨、全面。

在 G. Abramo 等 [44] 的研究中，使用 123128 篇 WoS 网站中的意大利出版文献进行研究，发现不同学科对预测模型的适用性不同。该研究对所有的文献进行研究主题分类，共分为“生物学”“生物医学”“化学”“临床医学”“地球与空间科学”“经济学”“工程学”“法律、政治和社会学”“数学”“交叉科学”“物理学”“心理学”12 个主题学科。其研究结果显示，“经济学”在两个预测模型中，早期引用具有最大的权重值，而“心理学”则相反；生命科学领域（“生物医学研究”“化学”“生物学”“临床医学”）的平均早期引用权重系数各不相同；“法律、政治和社会学”“工程学”和“交叉科学”都反映了明显的早期影响。

5.3 明确解决问题的前提，提出创新的预测路径

有时在解决实际问题的过程中，现实问题过于复杂，为了使复杂问题简单化，研究者会附加一定的前提条件，并在此前提条件下解决部分的问题。当去掉这个前提条件后，会出现什么样的问题，所研究的方法在实际操作中是否还能复现，值得我们思考和研究。

在使用动态异构信息网络对新出版论文进行引文时间序列预测的研究中 [58]，研究者认为以往的引文预测依赖论文发表后的头几年观察到的引文（即领先的引文价值），即通过头几年的被引量来预测长期的被引频次。然而现实情况是，许多论文在发表后的头几年其引用影响已经达到峰值，因此这些论文并未能体现出它的领先价值。在出版物更新频率非常快的领域（诸如机器学习）领域，等待 3-5 年才能预测影响是不现实的。基于此问题，该研究提出了一个挑战：为没有任何领先价值的新发表论文生成引文时间序列，解决时间序列任务中的“冷启动”问题。因此，他们提出了端到端的框架，即异构信息网络到时间序列，以此来预测单篇论文的被引频次。

该研究的核心思想是一种转化思想：通过学习由关键词、作者、出版地点和论文所构成的异构网络，估算出一个伪前导值，并将其映射为论文未来的引用时间序列，即将异构网络信息转换成时间序列信息，实现时间序列的预测。

5.4 运用数学建模思维提高模型的可解释性

由于前面所总结的基于经验主义的调参式机器学习、深度学习建模方法缺少数学工具去诊断和测评神经网络特征表达能力，缺乏可解释性，因此在这个建模过程中，可以根据自己研究的需求寻找合适的建模方法。数学建模思维是在现实情境中从数学视角出发，分析问题、提出问题、建立模型、确定参数、求解模型、并最终解决实际问题的一种思维方法。以下的建模方法充分地体现了建模过程的数学思维，并使用了数学工具对模型进行量化解释，充分地展示了模型的可解释性。

在论文的引文动力学机制研究中，M. Wang 等 [56] 从“论文引文模式能否预测长期影响”问题出发，首先确定了驱动论文被引用的三个基本机制：高被引论文比低被引论文更有可能被再次引用；论文具有老化效应，每篇论文的新颖性最终都会消失；论文存在内在差异。结合这三个因素，推导出论文被引用的概率模型：

$\Pi_i(t) \sim \eta_i C_i^t P_i^t$，其中 η_i 解释了论文的内在差异，因为论文的内在差异如新颖性、重要性等取决于多种无形和主观的维度，该研究忽略了评估一篇论文内在价值的必要性，并将合适的 η_i 视为一篇论文在研究总样本中内在差异的综合衡量标准；C_i^t 是论文 i 在发表后 t 时获得的引用；P_i^t 是论文 i 在发表后 t 时的衰减率。论文累计总被引频次可通过微积分的方式求解出。

该研究的创新点在于将引用预测视为一种连续型概率问题，通过推导概率密度函数，求得概率分布，以此求出论文的未来引用。相比于机器学习和深度学习等数学建模方法，在模型准确度大致相同的情形下，该种建模方式可解释性更强。

6 总结与展望

综上所述，在大数据、人工智能的时代下，引用预测研究内容不断更新，产生了新的影响因素指标和预测方法。本文从“影响因素”到“研究对象”“研究方法”进行了系统梳理，并在前人的研究中，总结了目前引用预测研究存在的问题，并提出了相应的建议。

未来应该深入理论研究，加强影响因素指标和研究方法的合理运用，找到合理的研究周期，建立统一的评价系统，完善研究的理论基础，并且在完善的理论研究基础上，着力于解决实际问题，充分运用宏观的数学建模思维，落实微观的具体操作方法，运用转化的思想，将复杂的实际问题转化为多个简单的问题，并逐一进行解决，使得模型能在实际问题中得到充分的应用。

参考文献：

[1] BARABASI A L, SONG C, WANG D. Publishing: handful of papers dominates citation[J]. Nature, 2012, 491(7422): 40.

[2] DIDEGAH F, THELWALL M. Determinants of research citation impact in nanoscience and nanotechnology [J]. Journal of the American Society for Information Science and Technology, 2013, 64(5): 1055-1064.

[3] BUELA-CASAL G, ZYCH I. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals[J]. Psicothema, 2010, 22(2): 270-275.

[4] FU L D, ALIFERIS C F. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature[J]. Scientometrics, 2010, 85(1): 257-270.

[5] YAN Y, TIAN S, ZHANG J. The impact of a paper's new combinations and new components on its citation[J]. Scientometrics, 2019, 122(2): 895-913.

[6] CHAKRABORTY T, KUMAR S, GOYAL P, et al. Towards a stratified learning approach to predict future citation counts[C]//IEEE/ACM joint conference on digital libraries (Jcdl): IEEE, 2014: 351-360.

[7] 柴嘉琪，陈仕吉．论文新颖性测度研究综述 [J]. 农业图书情报学报，2020, 32(10): 56-61.

[8] ANTONIOU G A, ANTONIOU S A, GEORGAKARAKOS E I, et al. Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature[J]. Annals of vascular surgery, 2015, 29(2):286-292.

[9] 魏瑞斌．论文平均引用时差与被引频次相关性分析 [J]. 情报杂志，2018, 37(2): 135-141.

[10] ROTH C, WU J, LOZANO S. Assessing impact and quality from local dynamics of citation networks[J]. Journal of informetrics, 2013, 6(1): 111-120.

[11] BARNETT G A, FINK E L. Impact of the internet and scholar age distribution on academic citation age[J]. Journal of the American Society for Information Science and Technology, 2008, 59(4): 526-534.

[12] BORNMANN L, SCHIER H, MARX W, et al. What factors determine citation counts of publications in chemistry besides their quality? [J]. Journal of informetrics, 2012, 6(1): 11-18.

[13] BISCARO C, GIUPPONI C. Co-authorship and bibliographic coupling network effects on citations[J]. Plos one, 2014, 9(6): e99502.

[14] LEIMU R, KORICHEVA J. What determines the citation frequency of ecological papers? [J]. Trends in Ecology & evolution, 2005, 20(1): 28-32.

[15] JAMALI H R, NIKZAD M. Article title type and its relation with the number of downloads and citations [J]. Scientometrics, 2011, 88(2): 653-661.

[16] ROSTAMI F, MOHAMMADPOORASL A, HAJIZADEH

知识管理论坛，2022（3）:299-313
DOI: 10.13266/j.issn.2095-5472.2022.025

M. The effect of characteristics of title on citation rates of articles[J]. Scientometrics, 2014, 98(3): 2007-2010.

[17] MCCABE M J, SNYDER C M. Does online availability increase citations? theory and evidence from a panel of economics and business journals[J]. Review of economics and statistics, 2015, 97(1): 144–165.

[18] STREMERSCH S, CAMACHO N, VANNESTE S, et al. Unraveling scientific impact: citation types in marketing journals[J]. International journal of research in marketing, 2015, 32(1): 64-77.

[19] ZHANG X, XIE Q, SONG M. Measuring the impact of novelty, bibliometric, and academic-network factors on citation count using a neural network[J]. Journal of informetrics, 2021, 15(2):101-140.

[20] MONTEFUSCO A M, NASCIMENTO F P, SENNES L U, et al. Influence of international authorship on citations in Brazilian medical journals: a bibliometric analysis[J]. Scientometrics, 2019, 119(3):1487–1496.

[21] 魏瑞斌 . 论文平均引用时差与被引频次相关性分析 [J]. 情报杂志 , 2018, 37(2):135-141.

[22] BORNMANN L, DANIEL H-D. Selecting manuscripts for a high-impact journal through peer review: a citation analysis of communications that were accepted by Agewandte Chemie International Edition, or rejected but published elsewhere [J]. JASIST, 2008, 59(12): 1841-1852.

[23] SKILTON P F. Does the human capital of teams of natural science authors predict citation frequency? [J]. Scientometrics, 2009, 78(3): 525-542.

[24] GUILERA G, GóMEZ-BENITO J, HIDALGO M D. Citation analysis in research on differential item functioning [J]. Quality & quantity, 2009, 44(6): 1249-1255.

[25] AKSNES D W. Characteristics of highly cited papers [J]. Research evaluation, 2003, 12(3): 159-170.

[26] ONODERA N, YOSHIKANE F. Factors affecting citation rates of research articles [J]. Journal of the Association for Information Science and Technology, 2015, 66(4): 739-764.

[27] COLLET F, ROBERTSON D A, LUP D. When does brokerage matter? citation impact of research teams in an emerging academic field [J]. Strategic organization, 2014, 12(3): 157-179.

[28] YU T, YU G, LI P Y, et al. Citation impact prediction for scientific papers using stepwise regression analysis [J]. Scientometrics, 2014, 101(2): 1233-1252.

[29] AIN Q-U, RIAZ H, AFZAL M T. Evaluation of h-index and its citation intensity based variants in the field of mathematics [J]. Scientometrics, 2019, 119(1): 187-211.

[30] AMARA N, LANDRY R, HALILEM N. What can university administrators do to increase the publication and citation scores of their faculty members? [J]. Scientometrics, 2015, 103(2): 489–530.

[31] NOSEK B A, GRAHAM J, LINDNER N M, et al. Cumulative and career-stage citation impact of social-personality psychology programs and their members [J]. Personality & social psychology bulletin, 2010, 36(10): 1283-1300.

[32] BORSUK R M, BUDDEN A E, LEIMU R, et al. The influence of author gender, national language and number of authors on citation rate in Ecology [J]. Open ecology journal, 2009, 2(1): 25-28.

[33] PENG T-Q, ZHU J J H. Where you publish matters most: a multilevel analysis of factors affecting citations of internet studies [J]. Journal of the American Society for Information Science and Technology, 2012, 63(9): 1789-1803.

[34] VAN DER POL C B, MCINNES M D, PETRCICH W, et al. Is quality and completeness of reporting of systematic reviews and meta-analyses published in high impact radiology journals associated with citation rates? [J]. Plos one, 2015, 10(3): e0119892.

[35] ROLDAN-VALADEZ E, RIOS C. Alternative bibliometrics from impact factor improved the esteem of a journal in a 2-year-ahead annual-citation calculation: multivariate analysis of gastroenterology and hepatology journals [J]. European journal of gastroenterology & hepatology, 2015, 27(2): 115-122.

[36] ZHU X P, BAN Z J. Citation count prediction based on academic network features [C]// Proceedings 2018 IEEE 32nd international conference on advanced information networking and applications (Aina). New York: IEEE, 2018: 534-541.

[37] DING Y, JACOB E K, ZHANG Z X, et al. Perspectives on social tagging [J]. Journal of the American Society for Information Science and Technology, 2009, 60(12): 2388-2401.

[38] 孔玲 , 王效岳 , 于纯良 , 等 . 学术论文离被引有多远——基于影响因素与预测方法的文献述评 [J]. 情报资料工作 , 2019, 40(6): 63-72.

[39] YAN R, HUANG C, TANG J, et al. To better stand on the shoulder of giants[C]// BOUGHIDA K. Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries. New York: ACM,2012:51-60.
[40] BUTUN E, KAYA M. Predicting citation count of scientists as a link prediction problem [J]. IEEE transactions on cybernetics, 2020, 50(10): 4518-4529.
[41] 耿骞 , 景然 , 靳健 , 等 . 学术论文引用预测及影响因素分析 [J]. 图书情报工作 , 2018, 62(14): 29-40.
[42] RUAN X M, ZHU Y Y, LI J, et al. Predicting the citation counts of individual papers via a BP neural network [J]. Journal of informetrics, 2020, 14(3): 101039.
[43] LOKKER C, MCKIBBON K A, MCKINLAY R J, et al. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study [J]. BMJ, 2008, 336(7645): 655-657.
[44] ABRAMO G, D'ANGELO C A, FELICI G. Predicting publication long-term impact through a combination of early citations and journal impact factor [J]. Journal of informetrics, 2019, 13(1): 32-49.
[45] BORNMANN L, LEYDESDORFF L, WANG J. How to improve the prediction based on citation impact percentiles for years shortly after the publication date? [J]. Journal of informetrics, 2014, 8(1): 175-180.
[46] 程子轩 , 张向先 , 郭顺利 . 基于作者特征和期刊特征的学术论文被引频次预测模型构建与分析 [J]. 情报科学 , 2021, 39(3): 179-184,192.
[47] YAN R, TANG J, LIU X, et al. Citation count prediction: learning to estimate future citations for literature [C]// Proceedings of the 20th ACM international conference on information and knowledge management. Glasgow, Scotland: Association for Computing Machinery,2011: 1247–1252.
[48] CHEN J P, ZHANG C X. Predicting citation counts of papers [C]//Proceedings of 2015 IEEE 14th international conference on cognitive informatics & cognitive computing. New York: IEEE, 2015: 434-440.
[49] AFZAL M, PARK B J, HUSSAIN M, et al. Deep learning based biomedical literature classification using criteria of scientific rigor [J]. Electronics, 2020, 9(8): 9081253.
[50] ABRISHAMI A, ALIAKBARY S. Predicting citation counts based on deep neural network learning techniques [J]. Journal of informetrics, 2019, 13(2): 485-499.
[51] YUAN S, TANG J, ZHANG Y, et al. Modeling and predicting citation count via recurrent neural network with long short-term memory[EB/OL].[2022-02-09]. https://arxiv.org/abs/1811.02129.
[52] WEN J Q, WU L Y, CHAI J P. Paper citation count prediction based on recurrent neural network with gated recurrent unit [C]//Proceedings of 2020 IEEE 10th international conference on electronics information and emergency communication. New York: IEEE, 2020: 303-306.
[53] XU J, LI M, JIANG J, et al. Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network [J]. IEEE access, 2019, 7: 92248-92258.
[54] DONG Y, JOHNSON R A, CHAWLA N V. Will this paper increase your h-index? [C]//Proceedings of the eighth ACM international conference on Web search and data mining. New York: ACM, 2015: 149-158.
[55] IBANEZ A, LARRANAGA P, BIELZA C. Predicting citation count of Bioinformatics papers within four years of publication [J]. Bioinformatics, 2009, 25(24): 3303-3309.
[56] WANG M, YU G, YU D. Mining typical features for highly cited papers [J]. Scientometrics, 2011, 87(3): 695-706.
[57] MA A, LIU Y, XU X, et al. A deep-learning based citation count prediction model with paper metadata semantic features [J]. Scientometrics, 2021, 126: 6803–6823.
[58] JIANG S, KOCH B, SUN Y. HINTS: citation time series prediction for new publications via dynamic heterogeneous information network embedding [C]// Proceedings of the Web conference 2021. New York: ACM, 2021: 3158-3167.

作者贡献说明：

张素芳：框架指导，提出修改意见，论文校对及定稿；

刘慧敏：论文撰写，数据整理。

ChinaXiv合作期刊
知识管理论坛，2022（3）:299-313
DOI: 10.13266/j.issn.2095-5472.2022.025

A Review of Research on Influencing Factors and Prediction of Citation Frequency of a Single Paper

Zhang Sufang Liu Huimin
School of Economics and Management, South China Normal University, Guangzhou 511400

Abstract: [Purpose/Significance] Combing the relevant influencing factors of the citation frequency of a single paper and the research status of the prediction of the citation frequency, this paper provides a comprehensive and systematic cognitive framework from the perspective of the involvement of scientific researchers and scientific research institutions in such research. **[Method/Process]** Using the literature research method, through the systematic combing of the existing literature, this paper summarized the relevant contents and characteristics of the influencing factors, research objects and research methods of citation prediction, compared and analyzed different methods by means of list, and summarized the common problems and some innovative solutions of the existing research. **[Result/Conclusion]** In the process of systematic combing and summarizing, it is found that the causal relationship between influencing factors and prediction results is not clear, the research sample data is lack of diversity, the relationship between the applicability of research results and prediction cycle is not clear, and the interpretability of model evaluation is weak. Therefore, we should improve the follow-up research quality from the aspects of solving the preconditions of the problem, selecting targeted samples, improving the extraction methods of influencing factors, and using mathematical thinking mode for modeling.

Keywords: the prediction of the citation frequency influencing factors regression analysis machine learning deep learning